The effects of signal erosion and core genome reduction on the identification of diagnostic markers

Jason W. Sahl[a,b], Adam J. Vazquez[a], Carina M. Hall[a], Joseph D. Busch[a], Apichai Tuanyok[c], Mark Mayo[4], James M. Schupp[b], Madeline Lummis[a], Talima Pearson[a], Kenzie Shippy[a], Rebecca E. Colman[b], Christopher J. Allender[a], Vanessa Theobald[d], Derek S. Sarovich[d], Erin P. Price[d], Alex Hutcheson[e], Jonas Korlach[e], John J. LiPuma[f], Jason Ladner[g], Sean Lovett[g], Galina Koroleva[g], Gustavo Palacios[g], Direk Limmathurotsakul[h,i], Vanaporn Wuthiekanun[h], Gumphol Wongsuwan[h], Bart J. Currie[d], Paul Keim[a,b], David M. Wagner[a#]

[a]Center for Microbial Genetics and Genomics, Northern Arizona University, Flagstaff, AZ; [b]Translational Genomics Research Institute, Flagstaff, AZ; [c]Emerging Pathogens Institute, University of Florida, Gainesville, FL; [d]Global and Tropical Health Division, Menzies School of Health Research, Darwin, NT, Australia; [e]Pacific Biosciences; [f]Division of Pediatric Infectious Diseases, University of Michigan; [g]Center for Genome Sciences, USAMRIID, Fort Detrick, MD; [h]Mahidol-Oxford Tropical Medicine Research Unit; [i]Department of Tropical Hygiene, Faculty of Tropical Medicine, Mahidol University, Bangkok

Running head: Core genome decay and signal erosion effects on diagnostics

#Address correspondence to David Wagner, dave.wagner@nau.edu

Abstract: Whole genome sequence (WGS) data are commonly used to design diagnostics for the identification of bacterial pathogens. To do this effectively, genomics databases must be comprehensive to identify the strict core genome that is specific to the target pathogen. As additional genomes are analyzed, the core genome size is reduced and there is erosion of the target-specific regions due to commonality with related species, potentially resulting in the identification of false positives and/or false negatives.

Importance: A comparative analysis of 1,130 *Burkholderia* genomes identified unique markers for many named species, including the human pathogens *B. pseudomallei* and *B. mallei*. Due to core genome reduction and signature erosion, only 38 targets specific to *B. pseudomallei/mallei* were identified. By using only public genomes, a larger number of markers were identified due to undersampling and represent potential false positives. This analysis has implications for the design of diagnostics for other species where the genomic space of the target and/or closely related species are not well defined.

Introduction

Whole genome sequence (WGS) data are routinely used to develop DNA-based diagnostics for rapid and accurate identification of clinical pathogens (1, 2). Validating the specificity of diagnostic targets ensures that assays do not produce false positives (identifying a non-pathogen as a pathogen) or false negatives (not identifying a pathogen that is actually present). To avoid false positives and negatives, DNA-based diagnostics must be conserved across the target species and absent from non-target species.

Two critical issues arise during the process of identifying specific diagnostics from bacterial genomes. First, the number of genes in the core genome (i.e. genes present in every individual of a species) tends to reduce as the number of sequenced genomes increases (3, 4). Certain pathogens (e.g. *Yersinia pestis*) propagate clonally, are highly homogeneous, and show little variation in core genome size with additional sampling (5). In this case, the core genome size is not expected to drastically reduce as more genomes are analyzed. In contrast, the core genome size of *Burkholderia pseudomallei* reduces significantly with each new genome added (6). A second issue arises from genomes of related species, or “near neighbors”, that share core genes with the target species. In a process of signature erosion, this genomic overlap often increases as near neighbor genomes are added to the analysis and erodes the number of potential diagnostic targets. Unfortunately, near neighbors are often under-sampled (or not sampled at all) during the search for diagnostic targets, which hinders efforts to identify species-specific targets.

Burkholderia represents a model genus for the demonstration of core genome reduction and signal erosion. The *Burkholderia* genus contains a diverse set of species, including plant pathogens (7) and human pathogens, such as *B. pseudomallei*, the causative agent of melioidosis (8), and *B. mallei*, the causative agent of glanders (9). The *pseudomallei* group includes *B. pseudomallei*, *B. mallei*, *B. oklahomensis*, *B. thailandensis*, and the newly described *B. humptydooensis* (10). The *B. cepacia* complex (*Bcc*) is a diverse group within *Burkholderia* that is associated with opportunistic infections and is comprised of

at least 20 genomic species (11, 12). Most of the relationships between these species have been determined through gene marker analyses, such as the *recA* gene (13, 14) or multilocus sequence typing (15).

From a genomics perspective, *Burkholderia* whole genome sequencing efforts have focused on *B. pseudomallei* (16) and *B. mallei* (17). Recent studies have begun to sequence other *Burkholderia* spp., including members of the *Bcc* (18). However, large-scale, whole genome, phylogenetics-based studies that define the overall phylogenetic structure among *Burkholderia* species using high-resolution methods are currently lacking.

In this study, we extensively surveyed the environment in Australia, the United States, and Southeast Asia for *Burkholderia* spp. We sequenced a large collection of genomes to: 1) explore the genomic diversity of *Burkholderia* spp. that grow on Ashdown's agar; 2) identify specific diagnostic markers for *B. pseudomallei* and *B. mallei*, and; 3) understand the sampling effect of core genome size reduction and signal erosion on the selection of highly specific diagnostic targets.

Results

Whole genome sequencing of *Burkholderia* spp. In this study, we analyzed the whole genome sequences of 829 *Burkholderia* spp. that grow on Ashdown's agar (Table 1), a selective medium containing the aminoglycoside, gentamicin. These isolates were collected from diverse geographic locations in the United States, Thailand, and Australia (Supplemental Table 1). To understand the effects of core genome reduction and signature erosion on the identification of highly specific diagnostic targets, the genomes of 256 diverse *B. pseudomallei*/*mallei* strains were sequenced, assembled, and deposited in public databases (Supplemental Table 1); these genomes were combined with 160 *B. pseudomallei*/*mallei* genome assemblies already in public databases. Most of the genomes (n=779) in this study were sequenced on the Illumina platform, with 50

genomes also sequenced on the PacBio platform, which generated highly contiguous, and often finished, assemblies (Supplemental Table 1).

Core genome SNP phylogeny. To understand the phylogenetic structure of the *Burkholderia* genus, genomes sequenced in this study as well as GenBank reference genomes (Supplemental Table 2) were aligned against *B. pseudomallei* K96243 (19) with NUCmer (20) and SNPs were identified with NASP. The maximum likelihood phylogeny inferred from core, orthologous SNPs (n=105,877) demonstrated that all genomes sequenced in this study, with the exception of 1 *B. gladioli* genome, grouped in either the *Burkholderia cepacia* complex (*Bcc*) or the *B. pseudomallei* group (Figure 1). Based on the monophyletic nature and complexity of the latter clade, we propose to name it the *B. pseudomallei* complex (*Bpc*). Multiple additional *Burkholderia* genomes from GenBank were analyzed and were found to be more distantly related to these two groups. Our clade naming scheme is consistent with a recently published taxonomic scheme for *Burkholderia* (21). As such, they were not examined in detail in this study but were included for marker screening purposes ("Paraburkholderia" genomes in Supplemental Table 2).

The work performed in this study greatly expands the known genomic diversity of *Burkholderia*. For example, at the onset of this study, only two *B. ubonensis* genome assemblies were available in Genbank. This is likely due to the fact that most genome sequencing has focused on clinically-relevant organisms, whereas we sampled environmental as well as clinical isolates. This study adds the genomes of 254 *B. ubonensis* isolates, including three finished genomes (3 contigs) and three nearly finished genomes (4-5 contigs) (Supplemental Table 1). All of these genomes are publicly available and will help provide phylogenetic context for additional *Burkholderia* genomes that are sequenced, including from clinical isolates. We have also generated the first WGS sequences for other recently described species, such as *B. stagnalis* and *B. territorii* (12), including completed genomes, which will provide data for additional comparative studies.

Comparative genomics. Based on the topology of the core genome phylogeny (Figure 1), pan-genome statistics were calculated for each major clade (Table 2) using LS-BSR (22). The core genome of each primary clade was aligned against all surveyed genomes (n=1,130) to identify species or clade specific markers. A marker was determined to be clade specific if it had a BSR value > 0.8 in all target genomes and < 0.4 in all non-target genomes; although this definition is very conservative, it was used to identify discriminatory markers, regardless of genome assembly quality. The results demonstrate that species-specific markers were identified for most of the major clades (Table 2); a multi-FASTA of all species-specific coding regions is publically available (https://gist.github.com/jasonsahl/3e4132ca1d09b717fcc2). A screen of these species-specific markers against all genomes was visualized to demonstrate their specificity to each targeted clade (Figure 2). The stability of markers from clades with a limited number of representatives is unknown and will need to be validated with additional sequencing. Markers also were identified for the *B. cepacia* complex (*Bcc*) and the *B. pseudomallei* complex (*Bpc*), which can help to verify results obtained through diagnostic sequencing efforts.

Putative new species. Based on the phylogeny (Figure 1), five divergent clades were identified that may represent novel species (PS-1 through PS-5). We have generated completed or nearly completed genomes for at least one isolate from each of these clades (Supplemental Table 1). A BLASTN alignment of the extracted *recA* sequences against the GenBank nucleotide database failed to identify a close match to a named species for any of these clades. To demonstrate the differences between genomes in these putative species, one representative was compared against a genome of the nearest species, based on closest patristic distance, or tree path length distance, to the nearest monophyletic clade in the global phylogeny (Figure 1). For each pairwise comparison, the ANI and DDH values were calculated and tabulated (Table 3). The results demonstrate that many of the clades have ANI values < 95%

compared to the nearest reference genome based on its position in the phylogeny. Putative species (PS)-2, which is most closely related to *B. oklahomensis*, demonstrated ANI values on the border of the species threshold when compared to *B. oklahomensis* genomes. All of the genomes from PS-2 have been isolated from Australia whereas all *B. oklahomensis*, including the two publicly available genomes, have been isolated from the United States (Supplemental Table 1). This physical separation, combined with the borderline ANI values, may argue for separate species, but biochemical testing is required to bolster this separation and is currently ongoing.

Core genome size reduction with additional sampling and signal erosion with the inclusion of near neighbor genomes. In bacteria with highly plastic genomes, the inclusion of additional isolates can cause the core genome size to decrease (3). To demonstrate this effect in *Burkholderia*, we calculated pan-genome statistics on 416 *B. pseudomallei*/*mallei* genomes. The results demonstrate that as additional genomes are added to the analysis, the core genome size reduces to 1,684 CDSs; annotation of these CDSs is provided in Supplementary Table 3. This analysis includes genomes from *B. mallei*, which has undergone significant evolutionary decay (9), and isolates from a chronic *B. pseudomallei* infection that have also undergone substantial genome reduction over time due to long-term host adaptation (23). By inclusion of a diverse set of genomes, the minimum set of genes required by all *B. pseudomallei/mallei* can be identified. From randomly subsampling the 416 genomes at different genome levels, the sampling effect on the core genome size was visualized (Figure 3a).

In addition to core genome size reduction, the effect of including additional near neighbor genomes on accurate diagnostics was also investigated. The core genome from all *B. pseudomallei/mallei* genomes was aligned using LS-BSR against a randomly selected subset of near neighbor genomes ranging from 10 to 300 genomes, with each iteration performed 100 times. By the time that 300 near neighbor genomes are randomly selected, the number of *B. pseudomallei/mallei* markers converged on the same number that is obtained

using the entire set of 714 near neighbor genomes (Figure 3b). This result demonstrates that a significant number of near-neighbor genomes must be sequenced in order to identify a set of molecular markers that are highly discriminatory for a given clade.

When we considered the 416 *B. pseudomallei/mallei* genomes, a surprisingly small number of unique markers (n=38) were identified (Supplemental Table 3). Of these markers, one in particular (BPSS0060; encodes a hypothetical protein) only contained 3 polymorphisms across all of the diverse *B. pseudomallei/mallei* isolates in our study. This gene represents a highly specific diagnostic target under low selection for mutation. If only *B. pseudomallei* is considered, 22 conserved markers were identified (Supplemental Table 4) in *B. pseudomallei* that are missing from *B. mallei* and all other considered *Burkholderia* genomes (Figure 2).

If we only consider the publicly available genomes used in this study (n=298), the core genome size for *B. pseudomallei/mallei* is 2,570 CDSs. When this core genome was screened against other *Burkholderia* near-neighbor genomes available in GenBank (n=141), 63 markers were identified that were unique to *B. pseudomallei/mallei.* In contrast, if near-neighbor genomes sequenced in the study were also included (n=573), 51 markers were identified. By not including additional non-target reference genomes, thirteen of these markers would represent false positives in screening studies. If only target genomes in GenBank are considered, 25 false positives would be identified, demonstrating the need to include large numbers of target and non-target genomes.

Discussion

Accurate design of highly specific diagnostics is important for the detection of dangerous human pathogens in both environmental and clinical settings. Timely pathogen identification directly from clinical specimens could inform the early treatment of potentially deadly infections. However, as our study demonstrates, the genomic targets of molecular assays need to first be thoroughly validated to avoid false positives and false negatives, which can potentially confound

diagnostic tests and delay appropriate patient treatment. In this study, we highlight the importance of exploring the strict core genome size and signature erosion before designing diagnostic PCR targets. Our genome-based approach is applicable for other researchers who wish to develop diagnostic assays for other pathogens.

The effect of signature erosion and core genome size reduction was highlighted in the genus, *Burkholderia*. To characterize the genomic space within *B. pseudomallei/mallei*, as well as in closely related genomes, we sequenced 829 *Burkholderia* genomes from diverse locations. A large-scale comparative genomics analysis of these genomes demonstrated that specific molecular markers were identified for many of the major *Burkholderia* clades identified from the core genome SNP phylogeny (Figure 2). These unique coding regions were likely acquired horizontally, based on the lack of homology of these regions in other lineages within the genus (Figure 2). To demonstrate the need to sequence a large collection of genomes to identify specific diagnostic targets, a core genome reduction analysis was performed (Figure 3a). This analysis demonstrated that sequencing additional genomes causes the core genome size to decline. This analysis was performed by including a large number of draft genome assemblies, which may cause genomic elements to be either truncated, based on unresolvable repeats, or missing altogether, based on either insufficient coverage or assembly algorithms that remove either short contigs or regions of anomalous coverage. Based on the genome panel used in this analysis, including a large and diverse set of isolates is important to avoid selecting potential diagnostic targets that are susceptible to false negative results when screening either clinical or environmental samples. If only genomes in public databases were selected, multiple markers would be identified that represent potential false negatives.

The other important factor to consider when designing diagnostic markers is the effect of signature erosion that can be introduced due to the inclusion of close relatives to the clade of interest. If only genomes available in GenBank were included in the analysis, 63 markers were identified that appeared to be specific

to *B. pseudomallei/mallei*. However, if all non *pseudomallei/mallei* genomes were included in the analysis, only 51 *B. pseudomallei/mallei* markers were identified, demonstrating the impact of including a comprehensive set of genomes outside of the targeted species or clade. If all genomes from our study were included, only 38 *B. pseudomallei/mallei* specific markers were identified, which demonstrates the need to include diverse genomes from both the targeted clade as well as from clades closely related to the targeted clade.

This study both expands the known genomic diversity of the *Burkholderia* genus and also provides a framework for using genomic data to design highly specific diagnostic targets. For some species, near neighbor genomes are not available or difficult to isolate, which complicates the identification of these targets, and highlights the need for continued genome sequencing. The reported sampling effects on strict core genome size and signature erosion must be considered when interpreting surveillance results for human pathogens.

Materials and Methods

Isolate collection, DNA extraction, genome sequencing, assembly. *Burkholderia* isolates were collected from diverse global locations with a focus on highly endemic regions for *B. pseudomallei*, including northern Australia and northeastern Thailand (Supplemental Table 1). Isolates were collected by the Menzies School of Health and Research, Northern Arizona University, the University of Michigan, the James Cook University, Mahidol University, and the US Army Medical Research Unit (USAMRU). All final culture and DNA extraction procedures were performed at Northern Arizona University, and whole genome sequencing (WGS) was performed at the Translational Genomics Research Institute (TGen; Illumina) and the U.S. Army Medical Research Institute of Infectious Diseases (USAMRIID; PacBio).

Isolates initially grown on Ashdown's agar were streaked from a single purified colony to form a lawn and then stored at -80°C in Luria Bertani (LB) broth with 20% glycerol. Cultures were grown on LB agar plates and incubated at 37°C for

24-48 hours. High molecular weight DNA was extracted using the Qiagen® DNeasy Blood and Tissue Kit (catalog no. 69504; Valencia, CA) for whole genome sequencing on Illumina (Illumina, Inc.; San Diego, CA) and Pacific Biosciences (Menlo Park, CA) platforms. Using approximately 2.7µg of gDNA, libraries were prepared for Illumina whole genome sequencing as previously described (24).

DNA was sequenced on multiple platforms, including Illumina HiSeq 2000, Illumina MiSeq, and PacBio. Raw Illumina reads were assembled with SPAdes v3.5.0 (25) in conjunction with a pipeline developed to identify sequence contamination between multiplexed samples (https://github.com/jasonsahl/UGAP). Contigs that showed an anomalously low depth of coverage compared to other contigs from the same assembly, or those that aligned to other organisms multiplexed in the same lane, were manually removed. Genome assembly information is shown in Supplemental Table 1.

For PacBio assemblies, genomic DNA was sheared to 20kb average size using g-TUBEs (Covaris inc.). After DNA damage repair and end repair, hairpin adapters were ligated to form a SMRTbell template. ExoIII and ExoVII treatment was used to remove failed ligation products. Size selection was performed on the Blue Pippin system (Sage Sciences) using 0.75% dye-free agarose gel cassette, marker S1 and Hi-Pass protocol; low cut was set on 4000 bp. Final library assessment was obtained by Qubit dsDNA BR assay and Agilent 2100 Bioanalyzer DNA 12000 chip analyses. Annealing of sequencing primer and binding polymerase P4 to the SMRTbell template was performed according to the PacBio calculator. The polymerase-template complexes were bound to MagBeads, loaded onto SMRTcells at final concentration 180 pM, and sequenced with 180 min movies on the PacBio RS II instrument.

PacBio sequences were assembled *de novo* using the Hierarchical Genome Assembly Pipeline (HGAP) (26). Draft assemblies were checked for overlapping ends using Gepard (27) and BLAST (28). Overlapping ends are typical of long-read assemblies of circular chromosomes. Redundant end sequences were

trimmed to one copy and the genome was rotated to create a new breakpoint. Reads were then re-aligned to the trimmed and shifted draft assembly for correction using the Quiver algorithm. Contigs that did not have identifiable homologous ends were corrected using Quiver without further processing.

Species identification using core genome single nucleotide polymorphism (SNP) phylogeny, average nucleotide identity (ANI), and digital DNA DNA hybridization (DDH) calculation. To model the evolutionary relationships between *Burkholderia* spp., a set of reference genomes (Supplemental Table 2) was downloaded from GenBank (29) and combined with the genomes sequenced in this study. For a number of these genomes, only raw reads were available, which were assembled for use in the comparative analyses described below. All genomes were aligned against the reference genome of *B. pseudomallei* K96243 (19) using NUCmer (20). Regions that aligned more than once by a reference self-alignment (i.e. duplicated regions) were removed from downstream analyses. All SNP-based methods were wrapped by the Northern Arizona SNP Pipeline (NASP) (http://tgennorth.github.io/NASP/) (30). Orthologous SNPs conserved in all genomes were concatenated and a maximum likelihood phylogeny was inferred with RAxML v8 (31) using the ASC_GTRGAMMA substitution model and Lewis correction (32).

For determining species differences, the average nucleotide identity (ANI) was calculated with default values in JSpecies (33). JSpecies calculates ANIb, which uses BLASTN alignments (28), or ANIm, which uses NUCmer alignments. The average values were reported over the entire length of all alignments. To find the nearest neighbor to which to query target genomes, the closest patristic distances were chosen, as calculated by DendroPy (34). Digital DNA-DNA hybridization (DDH) values were calculated with a web service (ggdc.dsmz.de) (35) and the range of reported values was reported.

Identifying *B. pseudomallei* and *B. mallei* markers for diagnostics using comparative genomics and pan-genome analysis. Coding DNA sequences

(CDSs) were identified for each species with Prodigal (36) and were de-replicated with USEARCH (37). Each representative CDS was then aligned against each genome with BLAT (38) and the Blast Score Ratio (BSR) (39) value was calculated; these methods were all wrapped by the Large-Scale Blast Score Ratio (LS-BSR) pipeline (22). LS-BSR was performed for each species and the number of core CDSs (BSR value > 0.8 in all genomes) in each group was calculated; a BSR value of 0.8 is roughly equivalent to 80% protein identity over 100% of the length of the protein (3). These core CDSs from a given species or clade were then screened against all other genomes, and those genes with a BSR value < 0.4 in all other species were identified as suitable species diagnostic markers.

The pan-genome was calculated for each clade using LS-BSR in conjunction with BLAT. A CDS was determined to belong to the core genome if it had a BSR value > 0.8 in all genomes queried for a given species or clade of interest. Each core CDS was then screened against all genomes in the analysis with LS-BSR. A CDS was determined to be species specific if it was in the core genome of the species or clade of interest and missing or highly divergent (BSR < 0.4) in all other *Burkholderia* genomes.

Core genome size reduction and signal erosion. To understand the sampling effect on the core genome size in *B. pseudomallei/mallei*, a set of 416 *B. pseudomallei/mallei* genomes was sampled without replacement from 1 to 400, with 100 iterations at each level. From each sub-sampling, a set number of genomes were randomly selected with a python script (https://gist.github.com/990d2c56c23bb5c2909d.git) and the core genome (CDSs with a BSR value of > 0.8 in all genomes) was calculated and plotted. *B. pseudomallei* and *B. mallei* were treated as a single species for this and many of the subsequent analyses as *B. mallei* is recognized as an equine-adapted clone within *B. pseudomallei* (40).

To understand the erosion of *B. pseudomallei/mallei* specific targets with the inclusion of sequences from other *Burkholderia* spp., the core genome (n=1,684

CDSs) from a set of 416 *B. pseudomallei/mallei* genomes was used. All *Burkholderia* near neighbor genomes (n=714) were then randomly sampled without replacement at different levels from 1 to 300. The *B. pseudomallei/mallei* core genome was then aligned against these near neighbor genomes to identify core regions present in other *Burkholderia* species, and the number of CDSs with a BSR value < 0.4 in all near neighbor genomes, indicating missing genes, was calculated and plotted.

Data Availability. Sequence data was submitted to the Sequence Read Archive for each isolate. Furthermore, genome assemblies for all isolates were submitted to NCBI. Individual accession numbers are show in Supplemental Table 1 and all data is deposited under PRJNA285704 and PRJNA279182.

Funding information. This work was supported by contract HDTRA1-12-C-0066 from the Department of Defense Chemical and Biological Defense program through the Defense Threat Reduction Agency to Dr. Wagner. The work at US Army Medical Research Institute of Infectious Diseases was funded by the Joint Science and Technology Office through the Defense Threat Reduction Agency, project CB10246.

Acknowledgements. Opinions, interpretations, conclusions, and recommendations are those of the authors and do not necessarily reflect the official policy or position of the US Army, US Department of Defense, nor the US Government.

Figure legends

Figure 1. A core genome single nucleotide polymorphism (SNP) phylogeny of *Burkholderia* genomes. All SNPs were identified by aligning genome assemblies against the finished genome of *B. pseudomallei* K96243 (19) with NUCmer (20) and processed with the Northern Arizona SNP pipeline

(http://tgennorth.github.io/NASP/) (30). A maximum likelihood phylogeny was inferred on the concatenated SNP alignment with RAxML v8 (31) with 100 bootstrap replicates. Clades were collapsed with ARB (41). Putative novel species are named PS (putative species) and the clade number.

Figure 2. A core genome single nucleotide polymorphism (SNP) phylogeny associated with a heatmap of markers unique to specific clades. The core genome phylogeny was inferred with RAxML (31) on a concatenated SNP alignment produced by aligning 1130 genomes against the finished genome of *B. pseudomallei* K96243 (19) with NUCmer (20) in conjunction with NASP (http://tgennorth.github.io/NASP/). Coding regions unique to specific clades were aligned against all genomes with LS-BSR (22) and the heatmap was visualized with the interactive tree of life (42). The heatmap demonstrates the distribution of identified markers against all genomes screened in this study.

Figure 3. (**A**) Core genome reduction in *Burkholderia pseudomallei/mallei*. The core genome was calculated with the LS-BSR pipeline (22) on 416 genomes. For sub-sampling, genomes were randomly selected at different depths and the number of coding regions (CDSs) with a blast score ratio (BSR) (39) value > 0.8 in all genomes was calculated and plotted. For each sub-sampling level, 100 iterations were performed. The mean value at each level is shown in red and each replicate is shown in black. (**B**) The effect of signature erosion on the design of *B. pseudomallei/mallei* diagnostic markers. Genomes outside of the *B. pseudomallei/mallei* clade (n=714) were randomly selected at different depths. The core genome of 416 *B. pseudomallei/mallei* genomes was screened against non *pseudomallei/mallei* genomes with LS-BSR (22) and the number of markers with a BSR value < 0.4 in non *pseudomallei/mallei* genomes was calculated and plotted. One hundred independent replicates were processed at each sampling depth. The mean value at each level is shown in red and each replicate is shown in black.

463
464
465
466
467
468
469
470

Table 1.Summary of new genomes sequenced as part of this study

clade	#genomes
B. anthina	8
B. cenocepacia 1	1
B. cenocepacia 2	4
B. cepacia	78
B. diffusa	12
B. gladioli	1
B. humptydooensis	5
B. lata	2
B. latens	2
B. metallica	1
B. multivorans	14
B. oklahomensis	2
putative species 1	3
putative species 2	4
putative species 3	10
putative species 4	7
putative species 5	8
B. pseudomallei	256
B. pseudomultivorans	9
B. pyrrocinia	1
B. seminalis	2
B. stagnalis	67
B. thailandensis	8
B. territorii	33
B. ubonensis	254
B. vietnamiensis	37
total	829

471

472
473
474
475
476
477
478
479

Table 2. Core genome stats

Species/clade	core genome size	#genomes	#species/clade specific markers
ambifaria	5408	2	71
anthina	5507	8	13
cenocepacia-1	3823	8	8
cenocepacia-2	5076	16	22
cepacia	4415	83	7
diffusa	4566	12	7
dolosa	5451	3	436
gladioli	4898	6	833
glumae	3253	3	264
humptydooensis	5115	7	157
lata	4214	7	0
latens	5348	3	105
multivorans	4001	21	53
oklahomensis	5681	4	141
PS-1	3693	3	504
PS-2	4231	4	23
PS-3	5047	11	195
PS-4	4366	7	0
PS-5	4978	8	0
pseudomallei	2339	392	22
pseudomallei/mallei	1690	416	38
pseudomultivorans	4549	10	62
pyrrocinia	6397	4	153
seminalis	6533	2	90
stagnalis	4835	67	54
thailandensis	4447	20	116

territorii	4399	33	0
ubonensis	3128	255	40
vietnamiensis	3803	40	71

480

Table 3. Average nucleotide identity (ANI) and DNA DNA hybridization (DDH) values between representatives of putative novel species and representatives of established clades

Genome	clade	nearest genome	ANIm (%)	ANIb (%)	DDH (%)
MSMB175	Putative species 1	*B. gladioli* BSR3	85.5	79.8	18.7-23.7
BDU8	Putative species 2	*B. oklahomensis* C6786	94.9	94.8	59.3-75.8
MSMB0852	Putative species 3	B. sp. MSMB43	92.4	91.1	44.5-52.7
MSMB0856	Putative species 4	*B. pyrrocinia* Lyc 2	91.2	89.8	44.9-60.8
NRF60-BP8	Putative species 5	*B. cenocepacia* KC-01	94.1	93.5	54.5-56.9

481

482

483

484

485

486

487

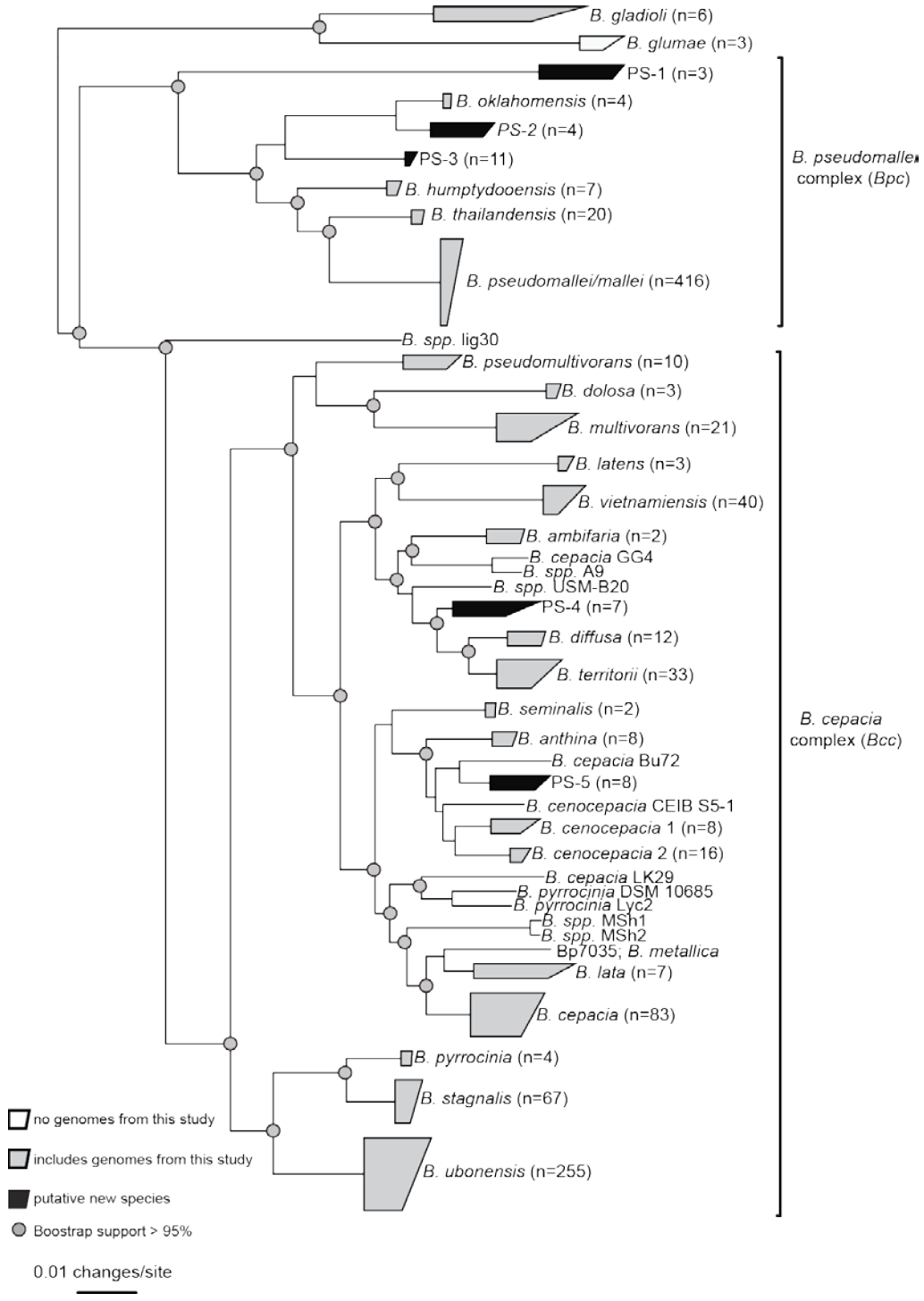
488

489

490

491

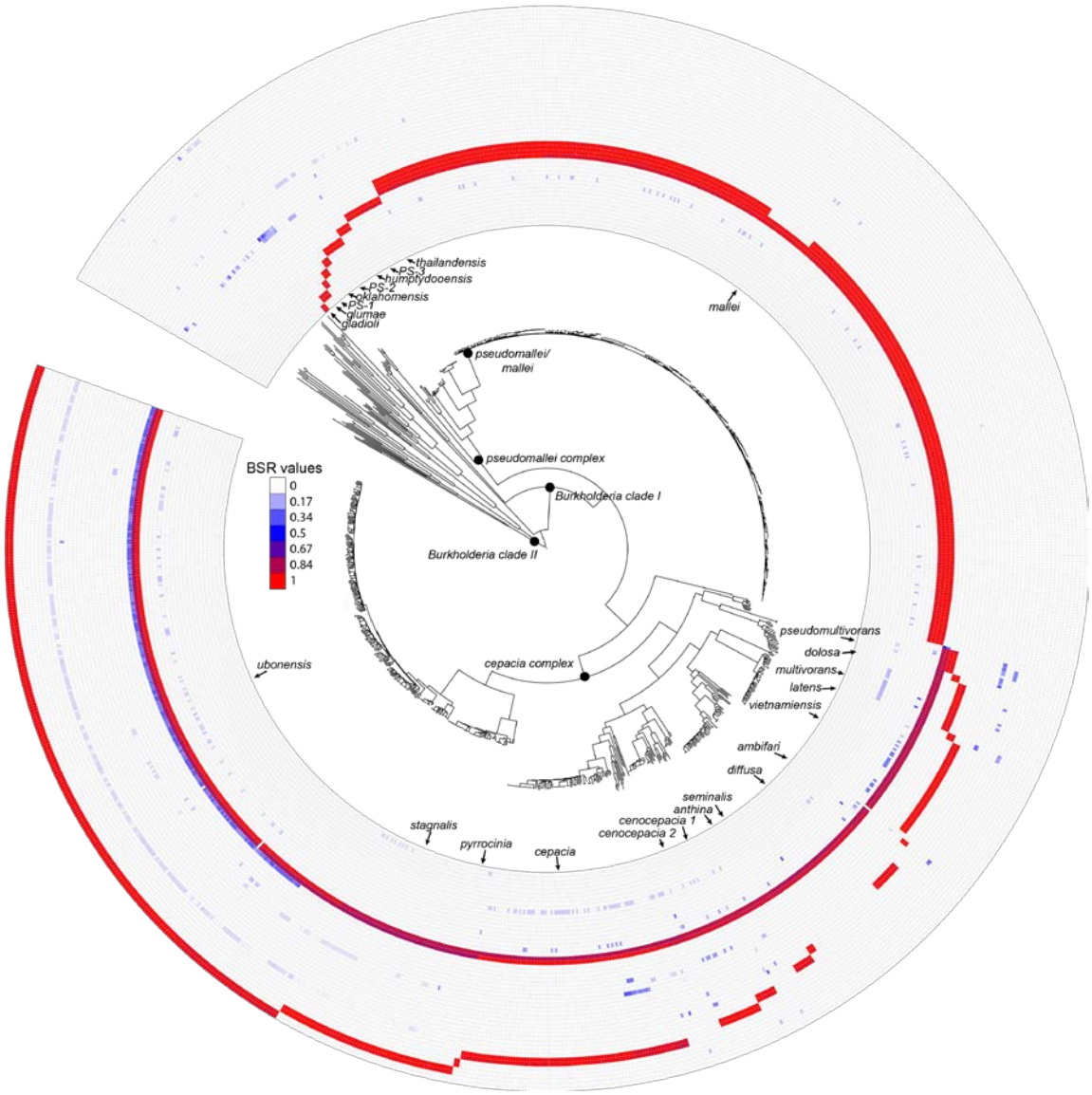
492 **Figure 1**



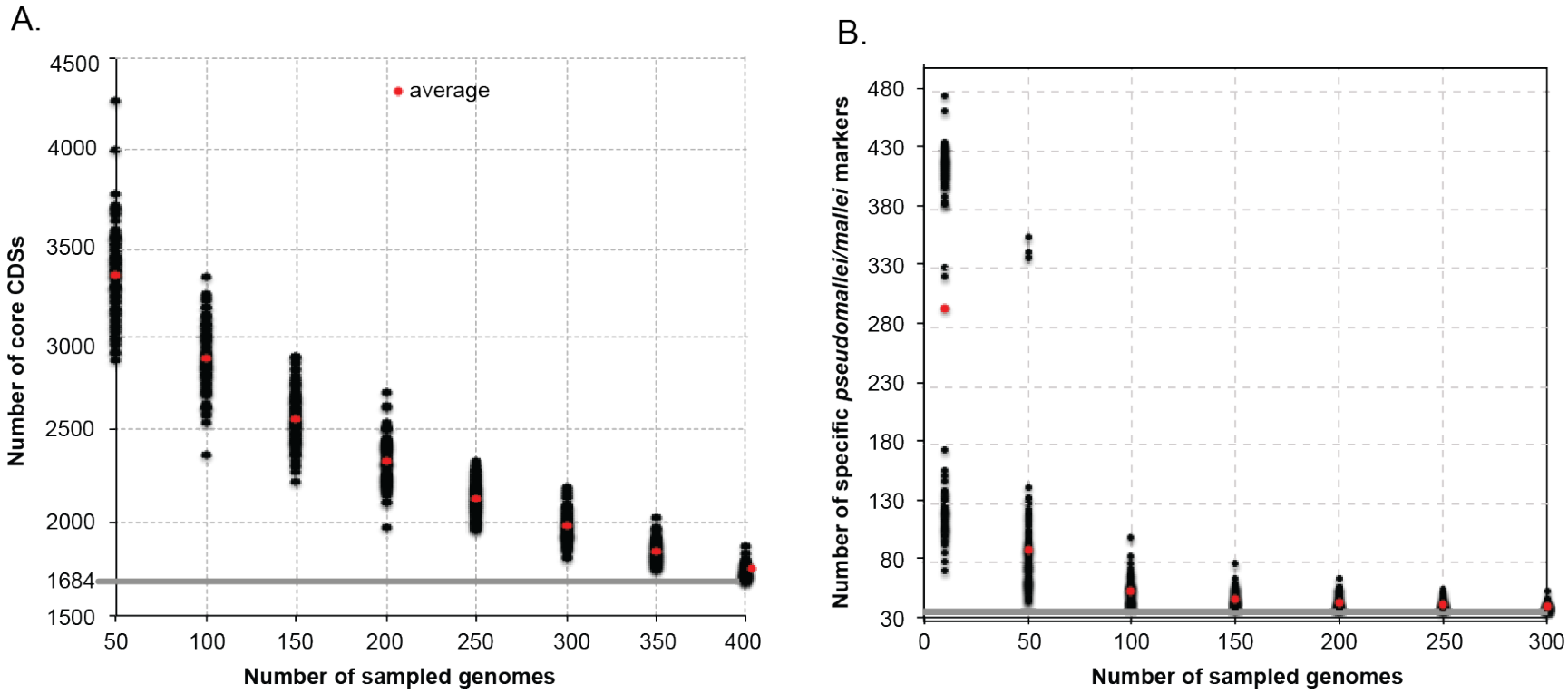
493

494

Figure 2


BSR values
0
0.17
0.34
0.5
0.67
0.84
1
thailandensis
PS-3
humptydooensis
PS-2
oklahomensis
PS-1
plumae
gladioli
mallei
pseudomallei/
mallei
pseudomallei complex
Burkholderia clade I
Burkholderia clade II
ubonensis
cepacia complex
pseudomultivorans
dolosa
multivorans
latens
vietnamiensis
ambifari
diffusa
seminalis
anthina
cenocepacia 1
cenocepacia 2
stagnalis
pyrrocinia
cepacia

506



507

508 Figure 3

509

510

511

512

513

References

1. **Driebe EM, Sahl JW, Roe C, Bowers JR, Schupp JM, Gillece JD, Kelley E, Price LB, Pearson TR, Hepp CM, Brzoska PM, Cummings CA, Furtado MR, Andersen PS, Stegger M, Engelthaler DM, Keim PS.** 2015. Using Whole Genome Analysis to Examine Recombination across Diverse Sequence Types of Staphylococcus aureus. PLoS One **10:**e0130955.
2. **Sahl JW, Morris CR, Emberger J, Fraser CM, Ochieng JB, Juma J, Fields B, Breiman RF, Gilmour M, Nataro JP, Rasko DA.** 2015. Defining the phylogenomics of Shigella species: a pathway to diagnostics. J Clin Microbiol **53:**951-960.
3. **Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R, Henderson IR, Sperandio V, Ravel J.** 2008. The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. J Bacteriol **190:**6881-6893.
4. **Califf KJ, Keim P, Wagner DM, Sahl JW.** 2015. Redefining the differences in gene content between Yersinia pestis and Yersina pseudotuberculosis using large-scale comparative genomics. Microbial Genomics **1**. doi:10.1099/mgen.0.000028.
5. **Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, Wagner DM, Feldkamp M, Kusecek B, Vogler AJ, Li Y, Cui Y, Thomson NR, Jombart T, Leblois R, Lichtner P, Rahalison L, Petersen JM, Balloux F, Keim P, Wirth T, Ravel J, Yang R, Carniel E, Achtman M.** 2010. *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. Nature Reviews Genetics **42:**1140-1143.
6. **Spring-Pearson SM, Stone JK, Doyle A, Allender CJ, Okinaka RT, Mayo M, Broomall SM, Hill JM, Karavis MA, Hubbard KS, Insalaco JM, McNew LA, Rosenzweig CN, Gibbons HS, Currie BJ, Wagner DM, Keim P, Tuanyok A.** 2015. Pangenome Analysis of Burkholderia pseudomallei: Genome Evolution Preserves Gene Order despite High Recombination Rates. PLoS One **10:**e0140274.
7. **Coenye T, Vandamme P.** 2003. Diversity and significance of Burkholderia species occupying diverse ecological niches. Environ Microbiol **5:**719-729.
8. **Wiersinga WJ, Currie BJ, Peacock SJ.** 2012. Melioidosis. N Engl J Med **367:**1035-1044.
9. **Losada L, Ronning CM, DeShazer D, Woods D, Fedorova N, Kim HS, Shabalina SA, Pearson TR, Brinkac L, Tan P, Nandi T, Crabtree J, Badger J, Beckstrom-Sternberg S, Saqib M, Schutzer SE, Keim P, Nierman WC.** 2010. Continuing evolution of Burkholderia mallei through genome reduction and large-scale rearrangements. Genome Biol Evol **2:**102-116.
10. **Gee JE, Allender CJ, Tuanyok A, Elrod MG, Hoffmaster AR.** 2014. Burkholderia pseudomallei type G in Western Hemisphere. Emerg Infect Dis **20:**682-684.
11. **Ho CC, Lau CC, Martelli P, Chan SY, Tse CW, Wu AK, Yuen KY, Lau SK, Woo PC.** 2011. Novel pan-genomic analysis approach in target selection for

multiplex PCR identification and detection of Burkholderia pseudomallei, Burkholderia thailandensis, and Burkholderia cepacia complex species: a proof-of-concept study. J Clin Microbiol **49:**814-821.

12. **Smet B, Mayo M, Peeters C, Zlosnik JE, Spilker T, Hird TJ, LiPuma JJ, Kidd TJ, Kaestli M, Ginther JL, Wagner DM, Keim P, Bell SC, Jacobs JA, Currie BJ, Vandamme PA.** 2015. Burkholderia stagnalis sp. nov. and Burkholderia territorii sp. nov., two novel Burkholderia cepacia complex species from environmental and human sources. Int J Syst Evol Microbiol doi:10.1099/ijs.0.000251.
13. **Payne GW, Vandamme P, Morgan SH, Lipuma JJ, Coenye T, Weightman AJ, Jones TH, Mahenthiralingam E.** 2005. Development of a recA gene-based identification approach for the entire Burkholderia genus. Appl Environ Microbiol **71:**3917-3927.
14. **Ginther JL, Mayo M, Warrington SD, Kaestli M, Mullins T, Wagner DM, Currie BJ, Tuanyok A, Keim P.** 2015. Identification of Burkholderia pseudomallei Near-Neighbor Species in the Northern Territory of Australia. PLoS Negl Trop Dis **9:**e0003892.
15. **Baldwin A, Mahenthiralingam E, Thickett KM, Honeybourne D, Maiden MC, Govan JR, Speert DP, Lipuma JJ, Vandamme P, Dowson CG.** 2005. Multilocus sequence typing scheme that provides both species and strain differentiation for the Burkholderia cepacia complex. J Clin Microbiol **43:**4665-4673.
16. **Nandi T, Holden MT, Didelot X, Mehershahi K, Boddey JA, Beacham I, Peak I, Harting J, Baybayan P, Guo Y, Wang S, How LC, Sim B, Essex-Lopresti A, Sarkar-Tyson M, Nelson M, Smither S, Ong C, Aw LT, Hoon CH, Michell S, Studholme DJ, Titball R, Chen SL, Parkhill J, Tan P.** 2015. Burkholderia pseudomallei sequencing identifies genomic clades with distinct recombination, accessory, and epigenetic profiles. Genome Res **25:**129-141.
17. **Nierman WC, DeShazer D, Kim HS, Tettelin H, Nelson KE, Feldblyum T, Ulrich RL, Ronning CM, Brinkac LM, Daugherty SC, Davidsen TD, Deboy RT, Dimitrov G, Dodson RJ, Durkin AS, Gwinn ML, Haft DH, Khouri H, Kolonay JF, Madupu R, Mohammoud Y, Nelson WC, Radune D, Romero CM, Sarria S, Selengut J, Shamblin C, Sullivan SA, White O, Yu Y, Zafar N, Zhou L, Fraser CM.** 2004. Structural flexibility in the Burkholderia mallei genome. Proc Natl Acad Sci U S A **101:**14246-14251.
18. **Johnson SL, Baker AL, Chain PS, Currie BJ, Daligault HE, Davenport KW, Davis CB, Inglis TJ, Kaestli M, Koren S, Mayo M, Merritt AJ, Price EP, Sarovich DS, Warner J, Rosovitz MJ.** 2015. Whole-Genome Sequences of 80 Environmental and Clinical Isolates of Burkholderia pseudomallei. Genome Announc **3(1) e01282-14**.
19. **Holden MT, Titball RW, Peacock SJ, Cerdeno-Tarraga AM, Atkins T, Crossman LC, Pitt T, Churcher C, Mungall K, Bentley SD, Sebaihia M, Thomson NR, Bason N, Beacham IR, Brooks K, Brown KA, Brown NF, Challis GL, Cherevach I, Chillingworth T, Cronin A, Crossett B, Davis P, DeShazer D, Feltwell T, Fraser A, Hance Z, Hauser H, Holroyd S, Jagels K,**

Keith KE, Maddison M, Moule S, Price C, Quail MA, Rabbinowitsch E, Rutherford K, Sanders M, Simmonds M, Songsivilai S, Stevens K, Tumapa S, Vesaratchavest M, Whitehead S, Yeats C, Barrell BG, Oyston PC, Parkhill J. 2004. Genomic plasticity of the causative agent of melioidosis, Burkholderia pseudomallei. Proceedings of the National Academy of Sciences of the United States of America **101:**14240-14245.

20. **Delcher AL, Salzberg SL, Phillippy AM.** 2003. Using MUMmer to identify similar regions in large sequence sets. Curr Protoc Bioinformatics **Chapter 10:**Unit 10 13.

21. **Depoorter E, Bull MJ, Peeters C, Coenye T, Vandamme P, Mahenthiralingam E.** 2016. Burkholderia: an update on taxonomy and biotechnological potential as antibiotic producers. Appl Microbiol Biotechnol **100:**5215-5229.

22. **Sahl JW, Caporaso JG, Rasko DA, Keim P.** 2014. The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. PeerJ **2:e332**.

23. **Price EP, Sarovich DS, Mayo M, Tuanyok A, Drees KP, Kaestli M, Beckstrom-Sternberg SM, Babic-Sternberg JS, Kidd TJ, Bell SC, Keim P, Pearson T, Currie BJ.** 2013. Within-host evolution of Burkholderia pseudomallei over a twelve-year chronic carriage infection. MBio **4 (4):e00388-13**.

24. **Engelthaler DM, Hicks ND, Gillece JD, Roe CC, Schupp JM, Driebe EM, Gilgado F, Carriconde F, Trilles L, Firacative C, Ngamskulrungroj P, Castaneda E, Lazera Mdos S, Melhem MS, Perez-Bercoff A, Huttley G, Sorrell TC, Voelz K, May RC, Fisher MC, Thompson GR, 3rd, Lockhart SR, Keim P, Meyer W.** 2014. Cryptococcus gattii in North American Pacific Northwest: whole-population genome analysis provides insights into species evolution and dispersal. MBio **5:**e01464-01414.

25. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA.** 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol **19:**455-477.

26. **Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J.** 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods **10:**563-569.

27. **Krumsiek J, Arnold R, Rattei T.** 2007. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics **23:**1026-1028.

28. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. J Mol Biol **215:**403-410.

29. **Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW.** 2012. GenBank. Nucleic Acids Res **40:**D48-53.

30. **Sahl JW, Lemmer D, Travis J, Schupp J, Gillece J, Aziz M, Driebe E, Drees K, Hicks ND, Williamson C, Hepp C, Smith DE, Roe C, Engelthaler DM, Wagner DM, Keim P.** 2016. The Northern Arizona SNP Pipeline (NASP):

accurate, flexible, and rapid identification of SNPs in WGS datasets. bioRxiv doi:http://dx.doi.org/10.1101/037267.

31. **Stamatakis A.** 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics doi:10.1093/bioinformatics/btu033.

32. **Leache AD, Banbury BL, Felsenstein J, de Oca AN, Stamatakis A.** 2015. Short Tree, Long Tree, Right Tree, Wrong Tree: New Acquisition Bias Corrections for Inferring SNP Phylogenies. Syst Biol doi:10.1093/sysbio/syv053.

33. **Richter M, Rossello-Mora R.** 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A **106:**19126-19131.

34. **Sukumaran J, Holder MT.** 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics **26:**1569-1571.

35. **Meier-Kolthoff JP, Auch AF, Klenk HP, Goker M.** 2013. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics **14:**60.

36. **Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ.** 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics **11:**119.

37. **Edgar RC.** 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics **26:**2460-2461.

38. **Kent WJ.** 2002. BLAT--the BLAST-like alignment tool. Genome Res **12:**656-664.

39. **Rasko DA, Myers GS, Ravel J.** 2005. Visualization of comparative genomic analyses by BLAST score ratio. BMC Bioinformatics **6:**2; DOI:10,.1186/1471-2105-6-2.

40. **Pearson T, Giffard P, Beckstrom-Sternberg S, Auerbach R, Hornstra H, Tuanyok A, Price EP, Glass MB, Leadem B, Beckstrom-Sternberg JS, Allan GJ, Foster JT, Wagner DM, Okinaka RT, Sim SH, Pearson O, Wu Z, Chang J, Kaul R, Hoffmaster AR, Brettin TS, Robison RA, Mayo M, Gee JE, Tan P, Currie BJ, Keim P.** 2009. Phylogeographic reconstruction of a bacterial species with high levels of lateral gene transfer. BMC Biol **7:**78.

41. **Jobb G, Forster W, Brettske I, Gerber S, Ginhart AW, Gross O, Grumann S, Hermann S, Jost R, Konig A, Ludwig W, Liss T, Lussmann R, May M, Nonhoff B, Reichel B, Strehlow R, Stamatakis A, Stuckmann N, Vilbig A, Lenke M, Strunk O, Ludwig T, Bode A, Schleifer KH, Westram R, Richter L, Meier H, Yadhukumar, Buchner A, Lai T, Steppi S.** 2004. ARB: a software environment for sequence data. Nucleic Acids Res **32:**1363-1371.

42. **Letunic I, Bork P.** 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics **23:**127-128.